Published in final edited form as:
Soc Networks. 2017 July ; 50: 17–25. doi:10.1016/j.socnet.2017.02.001.

Analytic Strategies for Longitudinal Networks with Missing Data

Kayla de la Haye,
University of Southern California

Joshua Embree,
University of Florida

Marc Punkay,
University of Florida

Dorothy L. Espelage,
University of Florida

Joan S. Tucker, and
RAND Corporation

Harold D. Green Jr.
RAND Corporation

Abstract

Missing data are often problematic when analyzing complete longitudinal social network data. We review approaches for accommodating missing data when analyzing longitudinal network data with stochastic actor-based models. One common practice is to restrict analyses to participants observed at most or all time points, to achieve model convergence. We propose and evaluate an alternative, more inclusive approach to sub-setting and analyzing longitudinal network data, using data from a school friendship network observed at four waves (*N*=694). Compared to standard practices, our approach retained more information from partially observed participants, generated a more representative analytic sample, and led to less biased model estimates for this case study. The implications and potential applications for longitudinal network analysis are discussed.

Keywords

longitudinal analysis; social networks; stochastic actor-based models; SIENA; analytic sample; friendship network

Corresponding Author: Dr. Kayla de la Haye, Department of Preventive Medicine, University of Southern California, CA, U.S.A., delahaye@usc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

BACKGROUND

Complete social network studies attempt to attain a census, or at least a relatively complete assessment, of the relationships among a bounded population (Carrington et al., 2005; Wasserman and Faust, 1994). That is, researchers define a study population within a meaningful social or physical boundary (for example, all students in a school, all staff in an organization, or all residents of a town) and attempt to gather data about the attributes and relationships of *all* "members" of this social network. Research designs and protocols that facilitate the collection of complete network data are ideal, and critically important for obtaining valid, non-biased information about network members, their ties, and the types of dynamics that play out within a particular social system. However, in studies that require individuals' consent and direct participation to gather these data, missing information from eligible network members who *do not* participate in the study is common, and referred to as "unit" non-response (Huisman and Steglich, 2008). In longitudinal studies that track changes in complete social networks over time, non-response by network members is problematic as participants who initially participate may later drop out of the study. Additionally, the members of the network may not be static, and people may join or leave the network boundary as it has been defined by the researchers (i.e., *network churn*). For example, students might join or leave a school mid-way through a study. Longitudinal network studies, therefore, may include actors with complete and partial unit non-response over the multiple observation points (see Huisman and Steglich, 2008 for additional detail).

Statistical models for complete network data, such as exponential random graph models/*p** (ERGM) (Robins et al., 2007a, 2007b) and stochastic actor-based models (SABM) (Snijders et al., 2010, 2007), are sensitive to missing data (Koskinen et al., 2010; Koskinen and Snijders, 2013). This is because non-response leads to missing information on the actor *and their relationships* (see Figure 1), and relationships or "ties" in social networks are inherently interdependent. Missing tie variables are known to affect the local and global structural properties of networks (Burt, 1987; Costenbader and Valente, 2003; Kossinets, 2006). In particular, this work has shown that missing tie data can have negative effects on centrality and degree measures (Costenbader and Valente, 2003; Kossinets, 2006) and clustering coefficients (Kossinets, 2006). As an example, if a network member who was particularly popular or who acted as a bridge between disparate social groups did not participate in a survey wave, the overall interpretation of the network structure could change dramatically (e.g., if node 4 were missing from Figure 1). Additionally, if missing tie information is related to the *value of a tie* and thus properties of the network (e.g., if ties are more likely to be missing from actors that are less central or isolated in the network), tie values are systematically missing (i.e., missing not at random) and summaries of the properties of the network will be particularly biased (Huisman and Steglich, 2008).

In an effort to minimize the negative implications of missing network data, social network statisticians are working on new approaches for the treatment of missing data. Developments of model-based approaches to estimate missing network data for single (cross-sectional) observations of networks are underway within the ERGM framework (Handcock and Gile, 2010; Koskinen et al., 2010; Robins et al., 2004), and include work on Baysian approaches (Koskinen et al., 2010) and likelihood-based approaches (Handcock and Gile, 2010). ERGM software, including PNet and statnet, provide procedures for fitting ERGMs to cross-sectional data with missing values that implement some of these approaches.

However, dealing with missing tie information in longitudinal network models is a more complex problem, which has been explored in a smaller number of studies (e.g., Huisman and Steglich 2008; Koskinen and Snijders, 2007). Huisman and Steglich (2008) examined the impact of missing data in SABMs, and evaluated various solutions that include restricting data analyses to completely observed cases (i.e., only network members with complete data), and different approaches to imputing missing tie information for complete and partial participant non-response (note: addressing network composition change, due to member churn, is discussed in Huisman and Snijders, 2003). They conclude that “missing actors have a large effect on analyzing longitudinal network data” (p.307) and that there are substantial problems if missing data are simply ignored or data analyses are restricted to complete cases. These problems include not being able to fit a model (convergence problems) and biased parameter estimates. Missing data imputation has been found to help with model convergence, however sophisticated imputation is needed to avoid underestimating uncertainty levels and generating biased parameter estimates. Huisman and Steglich (2008) conclude that the imputation processes implemented in the RSiena package, the primary software that implements the SABM, is currently the best solution. Generally, in the default Method of Moments estimation procedure, the missing data treatment at T_1 is that missing ties are absent ($x_{ij}^{imp}=0$), which essentially imputes the modal tie value of zero given the typical sparseness of social networks. At T_2 missing data are imputed by a deterministic value for the first observation: they are not replaced but are imputed in the simulation phase of the model estimation whereby *all actors* (including those with missing ties at T_1 and/or T_2) are free to make (unobserved) micro-changes to their outgoing ties. Overall, the model estimation is based only on the observed ties at both time points, but non-respondent actors can still have an effect on the simulated evolution of the network structure by playing a role in all actors’ choices for tie formation and dissolution. Maximum-likelihood (ML) estimation approaches in RSiena present a new alternative for estimating models with missing data that may provide advantages, and research is needed to evaluate when this approach should be implemented (see the RSiena Manual for additional documentation: Ripley et al., 2016). However, currently, ML estimation does not accommodate “joiners” and “leavers” in the network, meaning that it cannot be used to model longitudinal network data that has “churn” in the network members, where actors may join or leave the bounded population defined as the network of interest over the observed time periods. Therefore, this approach does not yet provide a good alternative to handling missing data with inconsistent network structure where members change over time.

Huisman and Steglich, and the RSiena manual (Ripley et al., 2016) note that the standard RSiena imputation approach, using the default Method of Moments estimation procedure, works well when the proportion of missing data is modest (< 20%). For example, our own experience indicates that the internal imputation approach often “works” (i.e., models will converge) when the amount of missing data is < 15–20% overall. However, in some cases models do not converge. Issues with model convergence due to missing data seem to be proliferating, as social network researchers explore more diverse social communities and relationship dynamics, and thus are evaluating change in larger, and often less consistent social networks over multiple (>2) points in time. Despite attempts at defining neat, segmented, socially connected communities as a "social network", and tracking these complete social groups over multiple observations, this is rarely the norm. For example, a mere 5% attrition rate at each wave over 2 waves results in 10% missing overall (i.e., if 100 participants are recruited into the study and 5% (*n*=5) fail to respond at wave 1 and then a further 5% of the remaining 95 (*n* = 4.75) are lost to attrition at wave 2, 10% of the original sample (*n*= 10) are missing, but 5% attrition at each wave calculated over 5 waves (number of participants with non-response at w1=5, w2=4.75, w3=4.5, w4=4.3, w5=4.1) results in 23% (*n*=23) missing overall. Overall, the problem of missing data compromising model convergence (as well as potentially leading to biased model estimates) when modeling longitudinal networks is increasingly common.

When researchers are faced with this problem of missing data that compromises the convergence of RSiena models, there are a small number of commonly adopted solutions. The majority of researchers to date appear to restrict their analyses to an analytic sub-sample that is observed (e.g., completes surveys) *at most or all of the observation time points* to facilitate model convergence, and then they apply the internal RSiena imputation to deal with any additional missing data among this subset of network members (e.g., Burk et al., 2012; de la Haye et al., 2013; Schaefer et al., 2011; Sentse et al., 2013). Specifically, researchers (including ourselves) who observe large, complete social networks over multiple points in time may select a subset of participants for their analysis who were observed at the majority of (e.g., 3 of 4) data collection points (note: whether this model is ultimately a good fit for the data can then be assessed using several metrics within the RSiena program, such as Jaccard coefficients and goodness of fit statistics (see Ripley et al, 2016)). This is less restrictive than focusing solely on participants with *complete cases* (i.e., those with complete data at all observations), but will still increase the likelihood that the SABMs will converge. However, a consequence of this analytic sampling approach is that network members who are rarely or inconsistently observed are excluded from the analytic sample. Moreover, it is possible (and perhaps likely) that these individuals differ in a meaningful way compared to those who participate consistently (for example, students who engage in deviant behavior may be more likely to miss school and not complete study surveys), resulting in tie and behavior data that are *missing not at random* and potential biases in the model estimates (Huisman and Steglich, 2008). Overall, excluding partially observed participants from studies of longitudinal networks may have important consequences for the conclusions drawn in this research, even when sophisticated missing data imputation is applied. Thus, a systematic exploration of alternative solutions seems warranted.

Current Study: Comparing Alternative Analytic Sampling Approaches for Longitudinal Network Analysis

This study uses longitudinal data (4 waves over 3 years) on youth alcohol use and school-based friendship networks to illustrate the issues outlined above, and to comparing approaches for dealing with missing data and defining analytic samples for longitudinal SABMs, including a new alternative strategy that may be more "inclusive" under certain conditions. Additionally, we validate our findings with a small simulation study.

Longitudinal complete social network methods (in particular SABMs) are now commonly employed to investigate how social networks influence risky behaviors in youth (Veenstra et al., 2013; Veenstra and Dijkstra, 2011). These studies typically invite all students in a well-defined population (e.g., a school, classroom, grade cohort, or small town) to participate in the study, and consent is gained from some proportion of eligible participants who are then essentially re-defined as the network population being studied. This percentage is often highly dependent on whether active or passive parental consent is required by the Institutional Review Board. Then, data are collected from consented students at multiple time points, although some study participants may not provide data at all waves for a multitude a reasons, including opting out of a survey or deviant behavior such as skipping school (leading to attrition of the study participants), or because they change schools or move out of the neighborhood (leading to members of the network "leaving" the bounded study population). Excluding these participants from the analytic samples used to model peer network effects on youth *risk behaviors* may be especially problematic.

In this paper, we attempt to fit a longitudinal network model (SABMs) in RSiena to the original data (Snijders et al., 2010, 2007). Because this model will not converge, we describe and compare two strategies to generate analytic samples that differ in the extent to which they exclude network members with missing data. The first is a "Stringent" sampling approach that is commonly employed in the literature, and includes a subset of participants who are observed in the majority (i.e., in 3 of 4) of waves. The second is an alternative "Inclusive" sampling approach that is rarely (if ever) employed in the literature as a strategy to accommodate missing data, which includes a subset of participants who completed a survey at *any two consecutive waves,* for whom we have information on at least one time period of change. We compare the descriptive statistics of the Stringent and Inclusive samples to the full set of participants, and anticipate that the Inclusive approach will generate an analytic sample that is larger and more representative of the original study sample. Standard SABMs are then fit to the Stringent sample, while a multi-group SABM approach is applied to the Inclusive sample to combine data from the network transitions between each set of consecutive waves into one model. We then evaluate differences in the results of the SABMs fit to the two samples and discuss the implications for the substantive conclusions drawn in this research. Finally, we summarize the results of a small simulation study we conducted that provides additional insights into these two approaches, and the implications of their application in future longitudinal network studies.

METHOD

Design, Procedure, and Sample

Data come from the University of Illinois Bullying and Violence Study (Espelage and Stein, 2007), which includes a sample of over 1,230 students from three diverse Midwestern public middle schools. "Saturated samples" (needed for complete social network data) were developed by inviting all enrolled Grade 6 through Grade 8 students to participate in the study, via letters to and meetings involving students, parents, and school staff. A waiver of active consent was approved by the institutional review board and school district administrators, and a 95% participation rate was achieved.

Six trained research assistants, the primary researcher, and a faculty member collected data. At least two of these individuals administered surveys to classes ranging in size from 10 to 25 students. Students were first informed about the general nature of the investigation. Next, researchers made certain that students were sitting far enough from one another to ensure confidentiality. Students were then given survey packets and the survey was read aloud to them. It took students approximately 40 minutes on average to complete the survey.

The current analyses focus on school friendship networks and alcohol use among middle school students between wave 1 and wave 4 (spring 2008 through fall 2009). We also limit our analyses to one middle school. This resulted in a total possible sample of $N = 694$ consented students from this middle school who completed at least one survey (the Full sample; see Table 1 for demographics of this sample). Although the friendships and alcohol use behaviors of these students were observed at multiple waves, there was also missing data and "churn" in the student body between wave 1 and wave 4. Change in the composition of this friendship network was due to the Grade 7 cohort finishing middle school at wave 3, a new Grade 6 cohort entering the schools at wave 2, and high turnover of students who changed schools, left the school, or who did not complete the surveys because they were absent (often due to suspensions).

Measures

Friendship network—Friendships among school-mates were assessed by asking participants to list the first and last names of "the kids at school that you hang out with the most." Eight response spaces were provided (although the number of friends to nominate was not specified) and they were instructed not to list siblings or friends who did not go to their middle school. These data were used to generate friendship networks at each wave, represented as a directed, asymmetrical adjacency matrix where 1 = a unilateral friendship between participants i and j and 0 the absence of a friendship.

Alcohol use—Three items from the *Problem Behavior Frequency Scale* (Farrell et al., 1992) were used to measure alcohol intake over the past year for Wave 1 (and since the time of their last survey at Waves 2, 3, and 4). The scale asked respondents if they had drunk beer, wine or wine coolers, or liquor, and each item specified that drinking meant more than a sip or taste. Response options included *never, 1 or 2 times, 3 or 5 times, 6 or 9 times,* and *10 or more times.* In the complete sample, a Cronbach's alpha coefficient was .90 was found for each wave of data collection. Because of the infrequent consumption of alcohol use in this sample of middle school students, responses to these items were used to compute a variable for "any alcohol use" at each wave (1 = any alcohol use, 0 = no alcohol use).

Covariates—Participants reported on their gender (0 = female, 1 = male), their race and ethnicity, and their grade in school. Parent education level was measured as a categorical variable that represents the highest level of education attained by their mother or father, as reported by the participant. This was transformed into a single dichotomous variable where 1 = *at least one parent had some college, completed college, or completed graduate school*, and 0 = *less than some college*.

Analytic Strategy

Below we describe the approach used to define the Stringent and Inclusive analytic samples, and the approach to specifying models for each of these samples. However, first we will describe the overall modelling strategy applied in this study: SABMs for longitudinal social network data (Snijders et al., 2010, 2007). SABMs are implemented in the RSiena 4.0 program (Ripley et al., 2016), and model the evolution and interdependence of complete social networks and the behaviors and attributes of the individuals (actors) in these networks. Two parts of this model are estimated simultaneously: a network dynamics sub-model tests effects predicting changes to friendship ties, and a behavior dynamics sub-model tests effects predicting changes to the dependent behavior variable (i.e., alcohol use).

The overall approach of the SABM algorithm is to simulate changes in the network and actor behaviors between multiple (>1) discrete observed panels of data, with the network evolving in continuous time as a Markov process. Unobserved network change is assumed to occur in ministeps, where actors have the opportunity to update (change or not change) one tie or behavior at each step. Model parameters identify processes that motivate actors' decisions to change their network ties or behavior, and the rate at which they make these changes. Some examples of typical model parameters include structural effects on network change (e.g., tendencies for actors to reciprocate network ties, or to send ties to network members who already have many ties), and effects that involve actor attributes (e.g., the tendency for actors to send ties to network members with similar attributes to themselves). For these analyses, model parameters were estimated using a method of moments procedure (note: we could not employ Maximum Likelihood estimation due to the changing composition of the network), and estimates deemed significant if the *t*-ratio (estimate divided by the standard error) was greater than 1.96. Missing data were imputed internally in RSiena (Ripley et al., 2016), and the maximum outdegree for the simulated networks was set at 8, because of this limit on the friendship nominations in the questionnaire. These models are described in detail in several publications (Snijders et al., 2010, 2007; Veenstra et al., 2013).

Ideally, our data would have minimal missing information about the actors and their ties across the four observed study waves, and we would proceed with defining the analytic sample (and thus longitudinal network) as all participants who completed surveys *in any* of the 4 study waves. However, RSiena models fit to this complete analytic sample would not converge due to too much missing data, and so the alternative approaches described below were investigated.

'Stringent' sample and single-network RSiena analyses—The Stringent analytic sample was comprised of participants who completed surveys in 3 of the 4 waves, and thus represents consented students who participated *consistently* in the study. Participants who completed fewer than 3 of 4 surveys were excluded from being "members" of the network and all available data on their attributes and friend nominations (sent or received) were excluded in these analyses. This resulted in a Stringent sample with 313 network members, from a possible 694 consented students in the Full sample. Missing data in the Stringent sample were dealt with in two ways: (1) if participants with missing data at a given wave were enrolled in the school and were eligible to complete a survey (but did not, or not completely), their missing responses were coded as missing (NA, which is the code for missing data in R); or (2) if participants were not enrolled at the school at a given wave (they had not yet joined the school, or had left the school), and they were not nominated as a “school friend” by their school-mates at that wave, then the actor was coded as not being a member of the network at that wave using the “network composition change” files in RSiena (see Ripley et al., 2016 for more detail about how ‘non-members’ at a given wave are treated in the model estimation). Based on these definitions, 116 actors were coded as having “joined” the network between wave 1 and wave 2 (no actors ‘joined’ at later waves) whereas 86 actors were coded as having ‘left’ the network between wave 3 and wave 4 (no actors ‘left’ the network at earlier waves).

We applied the standard single-network RSiena model, as described in Snijders (2010) and the RSiena manual (Ripley et al., 2016) to the Stringent sample network. This approach requires that *one* set of *n* actors and the relationships among them (i.e., one network) is observed at multiple discrete time points. Change in each time period (i.e., change between each pair of sequential panels of data) is modeled together in one SABM, with model parameter estimates assumed to represent homogenous processes across all time periods. There is the option to test for time heterogeneity in model effects across the different time periods (using a timeDummy function), and to include interaction terms between specific time periods and specific parameters to account for time heterogeneous processes among this set of actors (Lospinoso et al., 2010). Thus, the single-network SABM was used to identify processes that predict network and alcohol use dynamics between wave 1 and wave 4, among the set of 313 actors in the Stringent sample.

‘Inclusive’ sample and multi-group RSiena analyses—In this paper, we are proposing another alternative approach to defining an “Inclusive” analytic sample for longitudinal network analysis in RSiena. The aim is to retain more consented participants with missing survey data in the analytic sample, and therefore utilize more available data on the friendship patterns and behaviors of students who participated infrequently. To do this we decompose the set of study participants in this school ($N = 694$) into sub-groups of students who provided survey data over *any two consecutive waves* (for whom we have some measure of network and behavior change), and define each of these sub-groups as separate social networks. Specifically, for each time period in the study (i.e., a transition between any two waves from wave 1 to wave 4), students who provided survey data at the beginning and end of the time period are defined as a sub-group, and are therefore actors in this sub-group social network (note: any missing data in a subgroup would be a result of item non-response, and so was coded as missing (NA)). With four waves of data, there are three time periods and thus three sub-group networks with different (although overlapping) sets of actors. Figure 2 illustrates how students observed at each wave are segregated into the three sub-group social networks. This resulted in an Inclusive analytic sample of 427 (of 694 in the Full sample) who were included in *at least one* sub-group network.

The network dynamics of each sub-group network, which span a single time transition (e.g., from wave 1 to wave 2), can then be modeled collectively using a “multi-group” SABM in RSiena. The multi-group option allows us to combine multiple networks--or in this case multiple sub-group networks-- into one *multi-group analysis*. Each sub-group must have the same dependent and independent variables, although the number of actors in each group can vary. The different networks are then combined into one project and one model is specified for all subgroups in the school: the different subgroups are considered to be unrelated (with regards to the simulation steps); however, the groups are assumed to have the same model specification (i.e., the values of model parameters are assumed to be the same across sub-groups). This multi-group SABM analysis is more commonly applied for other procedures (see Koskinen & Edling, 2012; Snijders, 2001), including assessing time heterogeneity in longitudinal network analysis (Lospinoso et al., 2010), although to our knowledge, this is rarely (if ever) employed as a strategy to accommodate missing data. Thus, this strategy can also be used to account for heterogeneity in effects *across* sub-groups by testing for interactions between sub-group dummy variables and each model effect, and including interaction terms in the model when necessary (Lospinoso et al., 2010). Importantly, the multi-group model and default ‘single network’ model in RSiena are essentially the same SABM but just accommodate different data structures, meaning that they would produce identical results if both approaches were used to model complete network data.

Model specification details—A basic model was specified that included factors known to predict youth friendships and alcohol use, and effects that are recommended to adequately model dynamics of complex social networks in RSiena. The same model specification (i.e., the same set of effects) was fit to the Stringent and Inclusive samples (a model fit to the Full Sample would not converge). A forward selection approach, described in Burk, Steglich, & Snijders (2007) and Snijders et al. (2010) was used to select model effects and avoid issues of collinearity.

Predictors of friendship network dynamics: Several effects were included as predictors of friendship choices. Associations between alcohol use and friendship choices were tested with three effects: “alcohol use alter” is an effect of peers’ alcohol use on them receiving an actor’s friendship nomination; “alcohol use ego” is an effect of actors’ alcohol use on their outgoing friendship nominations; and “same alcohol use” captures the extent to which friendships were established between peers with matching alcohol use (based on the binary dependent variable, where 1 = any alcohol use). The roles of gender, school grade, race/ethnicity and parent education on friendship choices were included using the same friendship selection effects (covariate ego, covariate alter, same covariate). Finally, endogenous network effects included tendencies for actors to reciprocate friendships (reciprocity), to befriend friends of friends (transitive triplets) and popular peers (indegree popularity), and other structural features that are often important to adequately explain complex network dynamics (3-cycles, indegree activity, outdegree activity; see Ripley (2016) for detailed descriptions of these effects).

Predictors of alcohol use: Effects included as predictors of alcohol use were actor-level covariates (gender, school grade, race/ethnicity, parent education) and network effects. The latter were two types of effects testing friend influence on actor alcohol use (average similarity and total similarity), with a positive effect indicating that actors’ alcohol intake became similar to the intake of their nominated friend(s)), and an effect of network indegree, with a positive effect indicating that actors with the most friend nominations were likely to adopt or maintain the highest levels of alcohol use. Linear and quadratic shape effects were included to model the overall distribution of scores.

RESULTS

Descriptive Statistics for the Full, Stringent, and Inclusive Samples

A substantial proportion of participants in the full sample are excluded when defining analytic samples for the longitudinal network models (Table 1): of the 694 participants who consented into the study and completed at least one survey, just 313 are retained in the Stringent sample, and 427 in the Inclusive sample. Although the Stringent and Inclusive analytic samples are typically representative of the student demographics (gender, race/ethnicity, parent education) captured in the Full sample, the analytic samples differ in their prevalence of alcohol use and network characteristics. Overall there is a trend for the analytic samples to include participants who were less likely to use alcohol, and who were more popular as friends*,* compared to the Full sample, although the Inclusive sample appears to be the most representative of the Full sample. Significant differences in the descriptive statistics presented in Table 1 were found between the Full Sample and Stringent sample, where the average indegree was significantly higher in Stringent sample at Wave 2 ($t = -2.0$, $p = .04$) and Wave 3 ($t = -2.0$, $p = .04$).

SABM Results for the Stringent Sample and Inclusive Sample

Predictors of friendship dynamics—SABMs were fit to the Stringent and Inclusive analytic samples and the two models identified the same significant structural effects and covariate effects that predicted friendship dynamics (although the size of the estimates sometimes varied) (Table 2a). Overall, students had a significant preference to befriend peers who had nominated them as a friend (positive reciprocity) and who were friends of their friends (positive transitive triplets), and to form local triadic friendship hierarchies (negative 3-cycle). The negative indegree activity parameter indicates a tendency for students with high indegrees to make *fewer* friend nominations, reducing the correlation between indegrees and outdegrees. The positive outdegree activity parameter indicates a tendency for actors with high outdegrees to send (relatively) more outgoing ties. In addition to these structural effects, there was a significant tendency for males to receive more friend nominations (positive male alter), despite making fewer nominations themselves (negative male ego), and a significant tendency for students to befriend peers if they were similar to themselves on gender (positive same male), race (positive same race) and school grade (positive same grade).

Although the structural and covariate predictors were consistent in the Stringent and Inclusive sample models, the effects of alcohol use on friendship dynamics differed (Table 2a). In the Stringent sample only the “alcohol use ego” effect was significant, with the negative estimate indicating that alcohol users made fewer friend nominations. In this sample, alcohol use was not significantly related to the tendency to receive friend nominations (alter effect), nor was similarity in alcohol use a significant predictor of friendships. However, in the Inclusive sample there were three significant effects of alcohol use: alcohol users made fewer friend nominations (the same negative alcohol use ego effect found in the Stringent sample); alcohol users were less likely than non-users to receive friend nominations (negative alcohol use alter); and students preferred to befriend peers whose alcohol use differed to their own use (negative same alcohol use effect).

Predictors of alcohol use dynamics—In the Stringent sample and Inclusive sample models, there were no significant covariate or network effects found to predict change in alcohol use (Table 2b). The non-significant linear shape effect indicates that there was not significant change in alcohol use throughout these three years of middle school.

SIMULATION STUDY

An additional simulation study was conducted to evaluate the "Stringent" and "Inclusive" strategies for longitudinal network data analysis in RSiena. For this study, we used longitudinal data from the "Teenage Friends and Lifestyle study" (Michell & West, 1996) that is publicly available on the RSiena website (https://www.stats.ox.ac.uk/~snijders/siena/), and that includes three waves of data (collected over a two-year period from 1995–1997) from 160 secondary students in Glasgow. These students reported on their friendship ties (they could name up to 6 friends), alcohol use, and demographics (among other measures). This simulation study utilizes the following data from each of the three waves: participant's friendship network, defined by the presence (or absence) of friendship ties to other students in the school sample; self-reported alcohol consumption (where 1 = *no use*, 2 = *once or twice a year*, 3 = *once a month*, 4 = *once a week*, and 5 = *more than once a week*); and self-reported gender (male, female) and monthly pocket (spending) money in British Pounds.

Analytic Strategy for the Simulation Study

We sought to replicate the analytic approach described earlier in this paper. First, a standard single network SABM was fit to the complete data set of 160 students to establish a "true" model. Next, we simulated data sets where 10% and 25% of randomly selected participants had missing data (10 data sets were generated with each level of missing participant data), meaning that over the 3 waves, 10% (or 25%) of respondents had all of their data (individual attributes and outgoing friendship ties) coded as missing (NA) to replicate a situation of non-response on a survey at a given wave. Next, the Stringent and Inclusive analytic samples were derived from these simulated data sets with missing node-level data: Stringent analytic samples were comprised of participants who were not missing survey data in any of the 3 waves (i.e., participants with no missing survey response in 3 of 3 waves), and the Inclusive analytic samples were comprised of subgroups of participants who were not missing survey data in any 2 consecutive waves (i.e., no missing survey data at Wave 1 and Wave 2, and/or no missing survey data at Wave 2 and Wave 3).

Models were specified using the same approach described earlier in the paper. Predictors of friendship network dynamics included effects of alcohol use (for this continuous variable we included: alcohol use alter, alcohol use alter squared, alcohol use ego, and similar alcohol use), effects of the gender and pocket money covariates (covariate alter, covariate ego, same/similar covariate), and network effects (reciprocity, transitive triplets, 3-cycles, indegree-popularity square root), and outdegree-activity square root). Predictors of alcohol use dynamics included network effects (average similarity), effects of covariates (male, pocket money), and linear and quadratic shape effects (see Ripley et al. (2016) for additional descriptions of these effects). The standard single-network RSiena model (Snijders, 2010; Ripley et al., 2016), where one set of *N* actors and the relationships among them (i.e., one network) are modeled across the 3 discrete time points, was applied to the Full sample with no missing and to the Full samples with 10% missing data (*n* = 10 simulated data sets) and 25% missing data (*n* = 10 simulated data sets). These standard single-network RSiena models were also fit to the Stringent analytic samples: we fit a model to each of the 10 Stringent samples generated from the simulated data with 10% missing, and each of the 10 Stringent samples generated from the simulated data with 25% missing. For the Inclusive analytic samples, the multi-group SABM in RSiena was again used to combine the subgroup networks (the W1–W2 group and W2–W3 group) into one multi-group analysis, with one multi-group model fit to each of the 10 Inclusive samples generated from the simulated data with 10% missing, and each of the 10 Inclusive samples generated from the simulated data with 25% missing.

Results

Descriptive statistics for the Full samples, Stringent samples, and Inclusive samples are summarized in the supplementary material (see Table S1). As expected the Inclusive sampling approach does a better job of producing analytic samples that retain a greater proportion of the original Full sample, compared to the Stringent sampling approach. Because we simulated data sets with missing at random, there are not meaningful differences in the average alcohol use or demographics across the different samples (data missing not at random have the potential to bias the analytic sample as we saw in our study reported above). However, because more of the Full sample is retained in the Inclusive samples (vs. Stringent samples) the average value of friendship indegree is closer to the “true” value of this statistic in the Inclusive samples compared to the Stringent samples.

Results of the RSiena models are summarized in the Supplementary material (Table S2 and Figure S1, a through e). Overall, the results indicate that when 10% of data are missing at random, and the standard single group RSiena model will converge well for the Full sample (which it is likely to do), this (unsurprisingly) produces model results that are closest to the “true” model compared to both the Stringent and Inclusive sampling approach. However, when the overall level of missing data is substantially greater (i.e., 25%) and a model fit to the full sample *will not* converge well, forcing researchers to work with an alternate analytic subset, then the Stringent and Inclusive analytic approaches can produce results that differ from the “true model”. Overall, although there is bias in the estimates for both strategies, the Inclusive analytic approach is often (but not always) more aligned with the “true” model than the “Stringent” approach (see Table S2, and Figure S1 d and e).

DISCUSSION

Efforts to obtain comprehensive and complete longitudinal network data must be prioritized to ensure that the findings and recommendations generated by this research have minimal bias and error. Nonetheless, non-response by participants and natural variation in longitudinal social network studies remain common and present a challenge for modeling data, particularly when networks are unstable and observed at multiple time points over long periods of time. This paper documents two alternative strategies for reducing the proportion of missing data in analytic sub-samples that will help achieve model convergence, but that may differ in the extent to which they bias the sample and results, particularly when data are not missing at random.

In this proof of concept study of a school-based adolescent friendship network tracked over three years of middle school, we observe substantial missing data due to partial non-response despite very high initial participation rates (95%). Standard SABMs for longitudinal networks and behavior (Snijders et al., 2010) would only converge if we excluded the majority of participants with missing data from the analysis (i.e., any participant that had missing data on more than 1 of 4 waves), in line with the experience of many other researchers (Burk et al., 2012; de la Haye et al., 2013; Schaefer et al., 2011; Sentse et al., 2013). However, this generated an analytic sample that was notably more pro-social than the original sample: specifically, the analytic sample (comprised of students who participated in at least 3 of 4 waves) tended to have lower rates of alcohol use--the risk behavior of focus in this study--and higher proportions of popular students, as indicated by higher average friendship nomination indegree, relative to the full sample.

We also evaluated an alternate Inclusive approach to defining and analyzing the friendship network and alcohol use dynamics using multi-group SABM, which was able to retain a substantially greater proportion of the original sample in the analytic sample and identify new model effects. This approach defined the longitudinal friendship network data as multiple distinct networks that are each observed over a single time transition (a transition between any two sequential observations), retaining any participant that provided data over at least two consecutive waves. These networks were then analyzed using one multi-group network model (Ripley et al., 2016). This alternative Inclusive approach generated an analytic sample that was typically more representative of the students we originally observed, in terms of alcohol use and friendship indegree. The results of this SABM also identified three processes where alcohol use significantly influenced friendship network dynamics, whereas the SABM fit to the Stringent analytic sample of “consistent participants” only found evidence of one of these effects. Because the stringently defined analytic sample was particularly biased to excluding alcohol users and youth with fewer friends, the true variability in alcohol use and friendship nominations was reduced and the longitudinal relationships among these variables were no longer evident. The Inclusive sampling strategy and multi-group SABM would have preserved more of the natural variability in these variables because it retained a larger and more representative analytic sample, and because data imputation was unnecessary due to each “sub-group” network having complete data from participants at the start and end of the time transition. Overall, the proposed Inclusive sampling approach reduced the bias in the analytic sample, and so may have increased the validity of our findings. An additional simulation study, based on 10% and 25% missing data, provided additional support that when there is a high level of missing data (i.e., 25% across multiple waves) that causes difficulty in model convergence, the proposed Inclusive analytic sample and multi-group model approach produces SABM results that were *typically* closer to the “true” values of a model based on data with no missing, compared to the Stringent sampling and standard analytic approach. Nonetheless, it is important to note that *all* approaches to accommodating missing data did show some bias.

Despite the contribution of these findings to the field of longitudinal social network analysis, this study does have limitations that need to be recognized. This sampling and analytic strategy was developed and evaluated with a focus on analyzing social network dynamics using SABMs in RSiena, and therefore may not be applicable for when using other analytic strategies for longitudinal network analysis (e.g., relational event models, or temporal exponential random graph models). Also, the data that are the main focus of this study were a friendship network drawn from a study in one Midwest city, and so provide a case study for the implications of using different approaches to dealing with missing data, and possible implications in terms of SABM results. The simulation study results provide additional insights into these two approaches, and the implications of applying them to data missing at random, and missing not at random. Nonetheless, it will be important to replicate these findings with additional samples and types of social networks that may have different types and patterns of missing data.

Overall, this research outlines the challenges of missing data when modeling change in longitudinal networks, and documents important insights into approaches that can be adopted when models will not converge due to missing information. These approaches define analytic samples that remove network members with missing data, and have implications on the analytic sample and model outcomes. Our results clearly indicate that it is ideal to obtain data that has as little missing information as possible. However, when missing data compromises model convergence, researchers should carefully compare and select approaches for dealing with missing information. A strategy that we propose for defining and analyzing an “Inclusive” analytic sample is not commonly employed in practice, but may present a useful alternative for researchers: it may help to retain a greater proportion of network members with missing data, and thus may help to increase power and to detect real network effects, and avoid missing effects when data are not missing at random. The usefulness of these alternate approaches are likely to depend on the prevalence and patterns of missing information, and so researchers are recommended to explore multiple strategies to determine how they impacts their data and model estimates, as these decisions can impact their research findings in potentially profound ways.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Work on this article was support by grant R01DA033280-01 from the National Institute on Drug Abuse (PI: Harold D. Green). This research uses data from the University of Illinois Bullying and Violence Study (PI: Dorothy Espelage; grant from Centers for Disease Control Grant # 5 U49 CE001268-02). We would also like to thank Tom Snijders and Christian Steglich for their valuable advice on modeling large longitudinal social network data, and for their suggestions which led to the concepts being evaluated in this paper.

Abbreviations

ERGM	exponential random graph model
SABM	stochastic actor-based models

References

Burk WJ, Steglich CEG, Snijders TAB. Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. International Journal of Behavioral Development. 2007; 31:397–404. DOI: 10.1177/0165025407077762

Burk WJ, Van der Vorst H, Kerr M, Stattin H. Alcohol use and friendship dynamics: Selection and socialization in early-, middle-, and late-adolescent peer networks. Journal of Studies on Alcohol and Drugs. 2012; 73:89–98. [PubMed: 22152666]

Burt RS. A note on missing network data in the general social survey. Social Networks. 1987; 9:63–73. DOI: 10.1016/0378-8733(87)90018-9

Carrington, PJ., Scott, J., Wasserman, S. Models and Methods in Social Network Analysis, Structural Analysis in the Social Sciences. Cambridge University Press; New York, NY: 2005.

Costenbader E, Valente TW. The stability of centrality measures when networks are sampled. Social Networks. 2003; 25:283–307. DOI: 10.1016/S0378-8733(03)00012-1

de la Haye K, Green HD, Kennedy DP, Pollard MS, Tucker JS. Selection and influence mechanisms associated with marijuana initiation and use in adolescent friendship networks. Journal of Research on Adolescence. 2013; 23:474–486. DOI: 10.1111/jora.12018

Espelage, DL., Stein, N. Middle School Bullying and Sexual Violence: Etiological Models and Moderators. Centers for Disease Control; 2007. Grant # 5 U49 CE001268-02

Farrell AD, Danish SJ, Howard CW. Relationship between drug use and other problem behaviors in urban adolescents. Journal of Consulting and Clinical Psychology. 1992; 60:705–712. DOI: 10.1037/0022-006X.60.5.705 [PubMed: 1401386]

Handcock MS, Gile KJ. Modeling social networks from sampled data. Ann Appl Stat. 2010; 4:5–25. DOI: 10.1214/08-AOAS221 [PubMed: 26561513]

Huisman M, Snijders TAB. Statistical analysis of longitudinal network data with changing composition. Sociological Methods Research. 2003; 32:253–287. DOI: 10.1177/0049124103256096

Huisman M, Steglich C. Treatment of non-response in longitudinal network studies. Social Networks. 2008; 30:297–308. http://dx.doi.org/10.1016/j.socnet.2008.04.004.

Koskinen JH, Edling C. Modelling the evolution of a bipartite network—Peer referral in interlocking directorates. Social Networks. 2012; 34(3):309–322. http://dx.doi.org/10.1016/j.socnet.2010.03.001. [PubMed: 24944435]

Koskinen JH, Robins GL, Pattison PE. Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. Statistical Methodology. 2010; 7:366–384. DOI: 10.1016/j.stamet.2009.09.007

Koskinen JH, Snijders TAB. Bayesian inference for dynamic social network data. Journal of Statistical Planning and Inference, 5th St Petersburg Workshop on Simulation, Part II. 2007; 137:3930–3938. DOI: 10.1016/j.jspi.2007.04.011

Koskinen, JH., Snijders, TAB. Simulation, estimation, and goodness of fit. In: Lusher, D.Koskinen, JH., Robins, G., editors. Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications, Structural Analysis in the Social Sciences. Cambridge University Press; New York, N.Y: 2013.

Kossinets G. Effects of missing data in social networks. Social Networks. 2006; 28:247–268. DOI: 10.1016/j.socnet.2005.07.002

Lospinoso J, Schweinberger M, Snijders T, Ripley R. Assessing and accounting for time heterogeneity in stochastic actor oriented models. Advances in Data Analysis and Classification. 2010; :1–30. DOI: 10.1007/s11634-010-0076-1

Michell L, West P. Peer pressure to smoke: the meaning depends on the method. Health Education Research. 1996; 11:39–49.

Ripley, R., Snijders, TAB., Boda, Z., Voros, A., Preciado, P. Manual for RSiena. 2016. Available at: http://www.stats.ox.ac.uk/~snijders/siena/RSiena_Manual.pdf

Robins G, Pattison P, Kalish Y, Lusher D. An introduction to exponential random graph (p*) models for social networks. Social Networks. 2007a; 29:173–191. DOI: 10.1016/j.socnet.2006.08.002

Robins G, Pattison P, Woolcock J. Missing data in networks: exponential random graph (p*) models for networks with non-respondents. Social Networks. 2004; 26:257–283. DOI: 10.1016/j.socnet.2004.05.001

Robins G, Snijders T, Wang P, Handcock M, Pattison P. Recent developments in exponential random graph (p*) models for social networks. Social Networks. 2007b; 29:192–215. DOI: 10.1016/j.socnet.2006.08.003

Schaefer DR, Kornienko O, Fox AM. Misery does not love company. American Sociological Review. 2011; 76:764–785. DOI: 10.1177/0003122411420813

Sentse M, Dijkstra JK, Salmivalli C, Cillessen AHN. The dynamics of friendships and victimization in adolescence: A longitudinal social network perspective. Aggressive Behavior. 2013; 39:229–238. DOI: 10.1002/ab.21469 [PubMed: 23446945]

Snijders TA. The statistical evaluation of social network dynamics. Sociological Methodology. 2001; 31:361–395. DOI: 10.1111/0081-1750.00099

Snijders, TAB., Steglich, CEG., Schweinberger, M. Modeling the co-evolution of networks and behavior. In: van Montfort, K.Oud, H., Satorra, A., editors. Longitudinal Models in the Behavioral and Related Sciences. Erlbaum; Mahwah, NJ: 2007. p. 41-71.

Snijders TAB, van de Bunt GG, Steglich CEG. Introduction to stochastic actor-based models for network dynamics. Social Networks. 2010; 32:44–60. DOI: 10.1016/j.socnet.2009.02.004

Veenstra, R., Dijkstra, JK. Transformations in adolescent peer networks. In: Laursen, B., Collins, WA., editors. Relationship Pathways: From Adolescence to Young Adulthood. Sage; Los Angeles: 2011. p. 135-154.

Veenstra R, Dijkstra JK, Steglich C, Van Zalk M. Network-behavior dynamics. Journal of Research on Adolescence. 2013; 23:399–412. DOI: 10.1111/jora.12070

Wasserman, S., Faust, K. Social Network Analysis: Methods and Applications. Cambridge University Press; Cambridge: 1994.

HIGHLIGHTS

- Missing data in longitudinal social network studies are common and problematic
- We review approaches to longitudinal network analysis in RSiena with missing data
- Restricting the analytic sample to actors with complete cases is common practice
- An alternative approach can result in analytic samples that are more representative
- Differences in the results of RSiena models using these approaches are documented

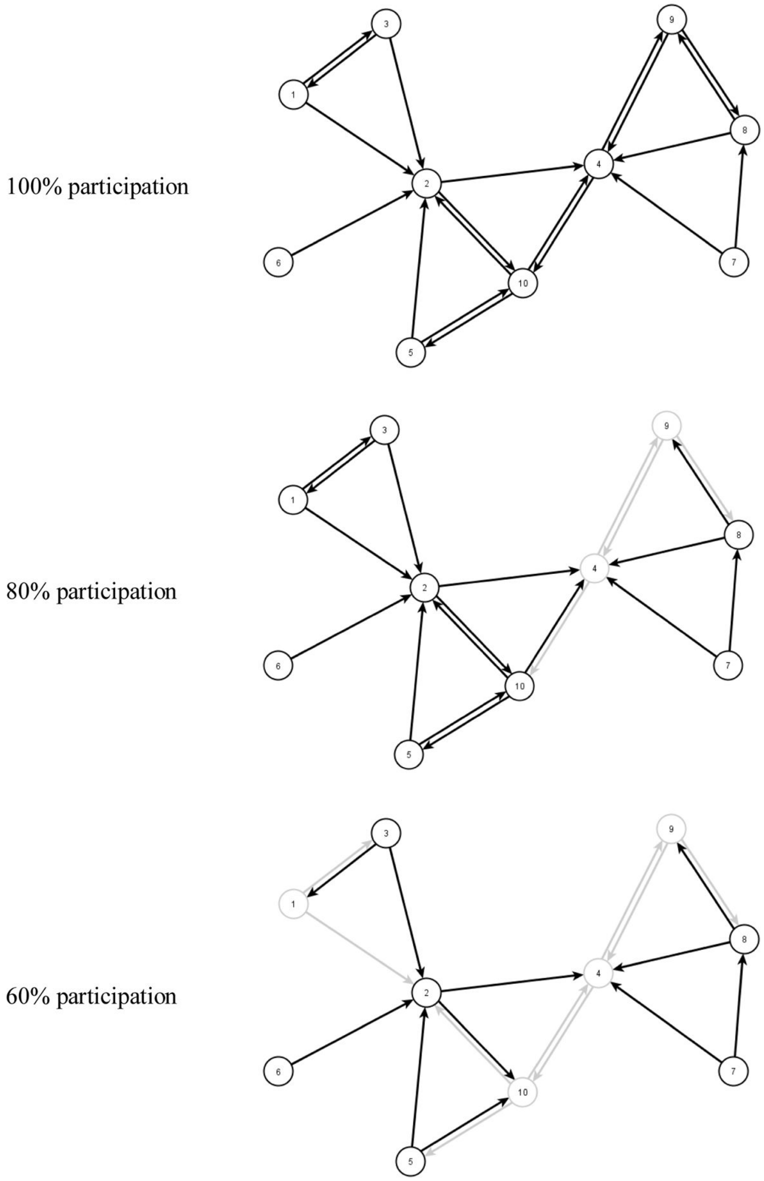

Figure 1.

Visualization of missing social network data. Nodes represent individuals, and directed lines represent relationships between a pair of individuals. With 80% participation, Node 4 and Node 9 are missing data (individual information, including their outgoing relationships). With 60% participation, Node 1, Node 4, Node 9 and Node 10 are missing data.

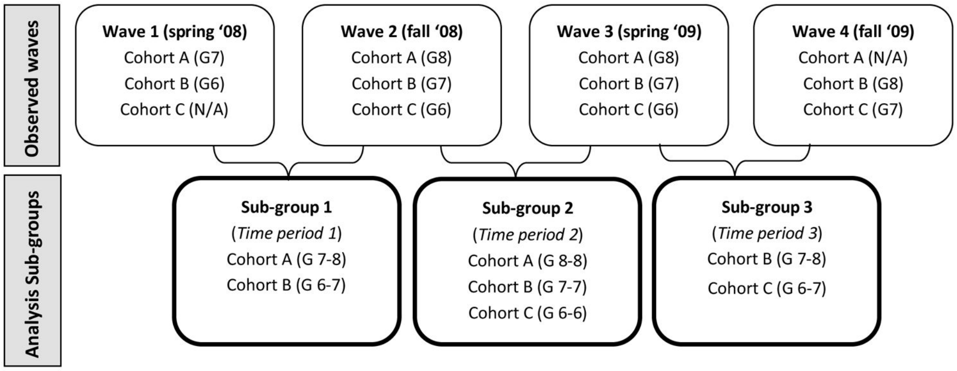

Figure 2.

Grade-level cohort observations across waves, and their inclusion in time period subgroups (G = grade; N/A = not observed at that wave)

Table 1

Demographic, alcohol use, and network characteristics of the full, stringent, and inclusive samples

Characteristic	Full Sample	Stringent Sample	Inclusive Sample
N	694	313	427
% male	51	50	50
Race/ethnicity (%)			
American Indian	2	2	1
African American	82	80	81
Asian	0	0	0
Hispanic	6	7	6
White	8	8	8
mixed	1	0	1
other	2	2	3
% parent with college education	62	61	61
Any past year alcohol use (%)			
W1	32.3	25.3	25.7
W2	30.1	26.2	29.5
W3	31.7	28.5	31.8
W4	29.6	29.9	29.6
M (*SD*) Friend indegree			
W1	4.5 (3.3)	4.8 (3.4)	4.7 (3.4)
W2	3.8 (3.4)	4.3 (3.3)	4.0 (3.4)
W3	3.6 (2.7)	4.0 (2.7)	3.7 (2.7)
W4	3.2 (2.7)	3.6 (2.7)	3.6 (2.7)

Table 2a

SABM results for Inclusive sample and Stringent sample: Effects on friendship dynamics

	Stringent Sample (Standard SABM)			Inclusive Sample (Multi-group SABM)			Difference
PARAMETER	**Est.**	***SE***	***p*-value**	**Est.**	***SE***	***p*-value**	
Effects on Friendships							
Rate period 1	26.55	2.44	0.000	17.30	1.82	0.000	
Rate period 2	10.96	0.75	0.000	11.35	0.71	0.000	
Rate period 3	25.77	4.38	0.000	24.35	2.72	0.000	
Structural effects							
outdegree	**−3.33**	0.12	0.000	**−3.19**	0.19	0.000	
reciprocity	**2.58**	0.09	0.000	**2.43**	0.09	0.000	
transitive triplets	**0.68**	0.05	0.000	**0.68**	0.05	0.000	
3-cycles	**−0.67**	0.08	0.000	**−0.61**	0.07	0.000	
indegree popularity (sqrt)	0.00	0.04	0.982	0.02	0.04	0.669	
indegree activity (sqrt)	**−0.70**	0.09	0.000	**−0.65**	0.08	0.000	
outdegree activity (sqrt)	**0.24**	0.03	0.000	**0.31**	0.03	0.000	
Covariates							
same race	**0.26**	0.06	0.000	**0.22**	0.05	0.000	
male alter	**0.14**	0.04	0.001	**0.16**	0.04	0.000	
male ego	**−0.11**	0.05	0.018	**−0.17**	0.05	0.000	
same male	**0.66**	0.04	0.000	**0.62**	0.04	0.000	
same grade	**0.43**	0.06	0.000	**0.64**	0.05	0.000	
parent education alter	−0.02	0.04	0.539	0.01	0.04	0.889	
parent education ego	−0.03	0.04	0.522	−0.04	0.04	0.416	
same parent education	0.02	0.04	0.608	0.06	0.04	0.119	
Any alcohol use							
alcohol use alter	−0.07	0.08	0.374	**−0.81**	0.31	0.010	*
alcohol use ego	**−0.21**	0.09	0.017	**−0.80**	0.30	0.008	
same alcohol use	N.S.			**−1.17**	0.36	0.001	*

Note. Est. = Unstandardized parameter estimate. N.S. = Not statistically significant. These effects were not included in the final model because they were found to be nonsignificant during the forward selection model specification. Effects in bold are significant at the p<.05 level.

Table 2b

SABM results for Inclusive sample and Stringent sample: Effects on alcohol use

	Stringent Sample (Standard SABM)			Inclusive Sample (Multi-group SABM)		
PARAMETER	**Est.**	***SE***	**p-value**	**Est**	***SE***	**p-value**
<u>Effects on Alcohol Use</u>						
Rate period 1	1.04	0.33	0.002	0.80	0.19	0.000
Rate period 2	0.86	0.18	0.000	0.65	0.12	0.000
Rate period 3	0.66	0.16	0.000	0.47	0.15	0.001
linear shape	−0.17	0.48	0.725	0.22	0.98	0.825
Covariates						
male	N.S.			N.S.		
grade	N.S.			N.S.		
parent education	N.S.			N.S.		
race/ethnicity						
Black	0.74	0.55	0.180	N.S.		
Hispanic	N.S.			−1.84	1.89	0.330
White	N.S.			N.S.		
Network influence						
average similarity	2.17	1.54	0.159	3.46	4.17	0.406
total similarity	N.S.			N.S.		
indegree	N.S.			N.S.		

Note. Est. = Unstandardized parameter estimate. N.S. = Not statistically significant. These effects were not included in the final model because they were found to be nonsignificant during the forward selection model specification. No differences in significant model effects were found.